Scores Assigned by Inexpert EFL Raters to Different Quality EFL Compositions, and the Raters' Decision-Making Behaviors

Turgay Han [i]
Ordu University

Abstract
The aim of this study is to examine the variability in and reliability of scores assigned to different quality EFL compositions by EFL instructors and their rating behaviors. Using a mixed research design, quantitative data were collected from EFL instructors' ratings of 30 compositions of three different qualities using a holistic scoring rubric. Qualitatively, think-aloud protocol data were collected concretely from a sub-sample of raters. The generalizability theory (G-theory) approach was used to analyze the quantitative data. The results showed that the raters mostly deviated while giving scores to very low level and mid-range compositions, but that they were more consistent while rating very high-level compositions. The reliability of the ratings of high quality papers (e.g. g: .87 and phi: .79 respectively) was higher than the coefficients obtained for mid-range and low quality compositions. This result indicated that more reliable ratings could be obtained in the rating of high quality papers. The think-aloud protocol analysis indicated that the raters attended differently to different aspects of these three level compositions. Implications are given from performance assessment practice perspectives.

Keywords: Inexpert raters, generalizability theory, variability of ratings, writing assessment.

[i] **Turgay Han** Department of English Language and Literature, Faculty of Science and Letters, Ordu University, 52200, Ordu, TURKEY.

Correspondence: turgayhan@yahoo.com.tr

Introduction

Performance assessment is a two-headed procedure that consists of real-life behavior observations or the simulation of that behavior, and in this sense assessing writing is performance assessment (Weigle, 2008) because it requires learners to show their actual writing performance. Assessing this performance is a difficult task due to the fact that the writing process has a multifaceted quality (Eryaman, 2008). The different aspects of writing performance make it challenging for raters to assess it. For example, social context may be a factor affecting the writing process (Baker, 2010). Further, the latter includes the language proficiency, conceptual knowledge and judgmental ability of the students (Heaton, 2003).

There are several arguments that suggest that assessing writing performance is a complex task. As writing is a complex part of language, assessing writing performance involves several variables. The first such variable is rater differences. While the rater's rating behavior is at one end of the spectrum (Gebril & Plakans, 2014; Lim, 2009), the rater's decision-making process is at the other end (Baker, 2010; Lim, 2009). Another issue is the rater's tendency to be severe or lenient (Huang, 2008; Lim, 2009). The rater's language background and rater training are also among the factors affecting the rating process (Chang, 2002; Shi 2001).

The effect of rubric type on writing scores is a second variable (Barkaoui, 2007; Han, 2013). Using different rating scales can contribute to scoring variance (Chang, 2002). While a holistic scale may be seen as suitable in some cases, an analytic scale may be favored in other circumstances (Bacha 2001; Knoch 2011).

The third and final variable is, inevitably, the learners. The learner's language proficiency (Huang, 2008;) and the effect of gender on writing scores (Green & Oxford 1995) are some factors which affect writing performance.

Beside the issue of the complexity of the writing assessment itself, the measurement of errors is among other issues associated with the assessment procedure (Brennan, 2011), because measuring the same trait more than once does not always give the same results, which raises the issue of the reliability of the measurement (Steyer, 2001). There are three theories which handle these issues, namely Classical Test Theory (CTT), Generalizability Theory (G-Theory) and IRT (Algina & Swaminathan, 2015; Brennan, 2010; 2011).

CTT is seen as the ancestor of G-Theory (Brennan, 2010; 2011). This theory is based on the equation that the observed score (X) equals the true score (T) plus random error (E) (Brennan, 2011). When a student is tested several times, the average of all the scores from these tests gives the true score in CTT (Rindskopf, 2015). Though this theory is frequently used in the social sciences, the most significant problem with the theory is its way of handling measurement error. While there are many factors affecting observed test scores, CTT brings together all the factors under the title of a single source of error (Brennan, 2010), and error is affected by investigator himself/herself even if s/he is not aware of this (Brennan, 2011). Because of this weakness of CTT, G-Theory was developed to deal with more than one error and with the extent to which various factors affect the errors (Matt & Sklar, 2015; Rindskopf, 2015). To give an example, imagine that each student is asked to write three short essays in a writing test and these essays are scored by two scorers; in this case, GT can be used to estimate the amount of variation which is caused by variation in essay topics and variation in raters (Rindskopf, 2015).

G-Theory is a theory that improves upon Classical Test Theory by investigating multiple errors and using analysis of variance (Brennan, 2010). Rather than reliability, dependability and generalizability are the terms used in G-Theory (Matt & Sklar, 2015). There are some strengths of G-Theory over Classical Test Theory. For example, G-Theory serves a wide range of areas, from education and business to medicine; similarly, it can be applied in a wide range of educational tests

and testing programs (Brennan, 2010). The conceptual framework that it provides is seen to be one of its most important strengths (Brennan, 2010). Despite its superiority over CTT, the use of G-Theory by researchers is relatively low, and it is thought that this reluctance to use G-Theory may be due to the "incomplete understanding of the conceptual underpinnings of GT, the actual steps involved in designing and implementing generalizability studies, or some combination of both issues" (Briesch, Swaminathan, Welsh & Chafouleas, 2014; p.13).

Overall, G-Theory enables the analysis of more than one measurement facet simultaneously in the assessment of error, reliability and variability in scores (Brennan, 2001). In this sense, this study used G-Theory as a methodological framework.

Given that there has not been much research into the scores assigned to different quality EFL compositions in a G-Theory framework with the use of think-aloud data to investigate rater behaviors, the present study investigated the variability in and reliability of the scores assigned by EFL instructors to different quality compositions within a G-Theory framework, and the raters' decision-making behaviors while rating the compositions and simultaneously thinking-aloud.

Literature Review

To date, there have been many studies on the factors affecting EFL/ESL writing scores as a result of rater impact, scale type impact and learner impact. Below is a review of the literature about these factors.

Rater Impact on Writing Scores

One dimension of EFL/ESL writing assessment research has addressed rater impact on writing scores (e.g. Lim, 2009; Shi, 2001). In the main, think-aloud protocols, interviews (Chang 2002; Gebril & Plakans, 2014), inventories (Alaei, Ahmadi & Zadeh, 2014) and case studies (Shi, 2001) have been used in the conducting of such research studies.

In this context, rater impact on writing and the rater's rating behaviors are among the notable facets. For example, Baker (2010) has found in a study that all raters have their own rating behaviors. Similarly, Lim (2009) argues in his doctoral dissertation that raters tend to give scores in two ways: using their own interpretation or judgment strategies. In addition, Gebril and Plakans (2014) have found that raters generally tend to use judgment strategies rather than interpretation strategies.

The rater's decision-making behavior is another important feature of writing assessment studies. In a study, Huang (2008) found that decision-making was a significant factor that could decrease the variations in the writing scores. Moreover, Lim (2009) addressed three types of decision-making process, including general impression, personal reaction and first impression. In addition, Baker (2010) reached the conclusion, in his study, that raters mostly had a tendency to give scores according to their first impression. Alaei, Ahmadi and Zadeh (2014) have also noted in their study that scores given according to first impression are more time-saving and cost-effective.

The rater's tendency to be severe or lenient also draws attention in the studies (e.g. Baker, 2010; Esfandiari & Myford, 2013; Huang, 2008; Lim, 2009). Esfandiari and Myford (2013) concluded from their study that teacher assessors were more severe than self-assessors and peer assessors, and that this situation gave rise to great variability in writing scores. On the contrary, Lim (2009) argued that differences in severity did not produce any significant variation in writing scores in the end. Furthermore, Huang (2008) has addressed this problem in terms of the numbers of raters and proposed that it is essential to prevent discrepancies in rater severity and leniency when a great number of raters participate in the assessing writing process.

The impact of raters' L1 background on writing scores has been studied. For instance, Chang (2002) found in a study that there was little difference between the scores of native and non-native English speaker raters. Similar to this result, Shi (2001) found that there was no noticeable difference in the scoring of native and non-native English speaker raters. However, Huang (2008) found that raters' language backgrounds had a significant effect on writing scores.

Several research studies have also examined whether rater training impacts upon writing scores (e.g. Alaei, Ahmadi & Zadeh, 2014; Chang, 2002). For example, Alaei, Ahmadi and Zadeh (2014) suggest in their study that rater training is an essential factor in making raters aware of their potential errors in the rating process. Similarly, Chang (2002) argues that rater training is necessary in order for raters to have the same or similar scoring philosophies and to provide inter / intra-rater reliability. Gebril and Plakans (2014) also suggest in their study that rater training is needed in order to purify and clearly articulate the scoring decisions. On the other hand, Esfandiari and Myford (2013) have argued that diversity in writing assessment scores could not be eliminated to a high extent with rater training. Recently, Han (2013) examined whether holistic scores could be as reliable as analytic scores when raters received detailed rater training. It was found that holistic scores were as reliable as analytic scores.

Scale Type Impact on Writing Scores

A second dimension of ESL writing assessment research has analyzed the impact of the rating scale on writing scores (e.g. Barkaoui, 2007; Chang, 2002; Han, 2013; Huang, 2008.). In his study, Chang (2002) found that there were significant scoring differences when raters used particular scale types. Similarly, Saeidi and Semiyari, (2011) suggested that there was a considerable difference between holistic and analytic scores. Barkaoui (2007) also noted this difference, but the findings of his study suggested that the holistic scale gave more reliable scoring results. In contrast, Knoch (2011) suggested that an analytic scale gave more reliable results and that this scale type had a significant effect on writing scores because of addressing a variety of descriptors differently from the holistic scales. Like Knoch (2011), Bacha (2001) favored analytic scoring in regard to its level of being informative compared with the holistic scale. Furthermore, Huang (2008) addressed the significance of the impact of rating scale types, and proposed that, unless rating scales provide a system for addressing the potential differences, they culminate in less consistent scores. In contrast to these findings, Alaei, Ahmadi and Zadeh (2014) found that some raters did not follow any holistic or analytic scale, but that they used their own rating styles which were not in accord with the criteria on rating scales. In addition, Reazaei and Lovorn (2010) also noted in their study that the raters had a tendency to give scores regarding the mechanical features of the students' writing, instead of the content, whichever scale they used.

Learner Impact on Writing Scores

A third dimension of study has addressed the scoring differences caused by learners (e.g. Baba, 2009; Huang, 2008; Lim 2009). The learners' proficiency levels in the EFL/ESL and language backgrounds are among the concerns of these studies. Huang (2008) suggested that ESL students had lower scores than Native English students; in addition, he argued that ESL students may have difficulty in understanding the writing tasks compared to Native English students and that they had difficulty while writing because of their linguistic deficiencies. Baba, (2009) has also reported a similar finding, asserting that the ability to use words appropriately to express ideas in a second language makes an excellent contribution to writing performance and scores. This finding is also supported by Kobrin, Deng and Shaw (2011).

There is some research which has studied the factor of gender. However, Lim 2009 has suggested that gender has little effect on writing scores, and that it is in favor of female learners. Green & Oxford (1995) have argued that women have a higher level of performance than men. Such research has focused on different aspects of writing, while examining the gender factor. For example, Breland, Bridgeman, and Fowles (1999) address this issue in terms of test types that have multiple-choice writing tests and essay writing, while Willingham and Cole (1997) address it in terms of prompts.

Quality of Papers

Very little research has examined the link between the quality of a paper (low, medium or high) and the reliability of and variability in the scores assigned by raters (e.g. Cumming, 1985; 1990; Huang et al., 2014).

Cumming (1985) found that three ESOL raters focused their attention on different qualities while rating the same composition. Next, Mendelsoh and Cumming (1987) examined the difference between the scores assigned to different level ESL compositions and the perceptions of 26 university professors from different academic disciplines (i.e., Engineering, English literature, ESL). The results showed that engineering professors agreed on the ratings of high and low quality papers but differed on the ratings of the middle quality papers. Further, engineering professors gave more importance to language features than to rhetoric and the organization of ideas, while ESL professors attended more to rhetoric and the organization of ideas.

Cumming (1990) investigated whether raters distinguish students' writing proficiency and language proficiency while rating compositions holistically and how the raters (6 experienced and 7 inexperienced) behaved in this decision-making process. The EFL/ESL teachers rated 12 compositions by students with different proficiency levels (intermediate and advanced) and writing expertise (average and professionally experienced writers) in their L1. The results showed that all teachers distinguished students' L2 proficiency and writing skills as separate, non-interacting factors. Further, the raters' concurrent verbal report analyses indicated that both groups of raters differed significantly in using most of the 28 common decision-making behaviors. For example, the inexperienced and experienced raters differed significantly in terms of their ratings for "content" and "rhetorical organization", but not for "language use". The rater reliability of inexperienced raters' ratings for content and rhetorical organization was higher than that of the expert raters. Additionally, both groups of raters' ratings for language use significantly differed from their ratings for content and rhetorical organization. Overall, the ratings of the experienced group of raters for the three components of writing were consistent.

In a TOEFL research Project, Cumming, Kantor and Powers (2001) developed a framework in respect of 10 experienced raters' decision-making processes while rating ESL/EFL writings. The data was collected through think-aloud protocols. The results showed that the ESL/EFL raters paid greater attention to rhetoric and ideas in the very high quality essays compared to very low quality essays. ESL/EFL raters paid more attention to rhetoric and ideas in high quality essays compared to very low quality essays, while they consistently attended to language features in the high quality essays compared to the very low quality essays. Native-English-composition raters behaved in a similar manner.

More recently, Huang et al. (2014) investigated whether the quality of essays (i.e., low quality vs. medium quality vs. high quality) interfered with the assessment of ESOL students' writing at a Turkish university, using G-Theory for analyzing the data. Five EFL raters rated 9 compositions (3 low, 3 medium and 3 high quality) by undergraduate level EFL students, both holistically and analytically. The G-Theory approach was used to analyze the data. Raters were found to be more consistent in rating high quality papers. Further, scoring methods greatly impacted on the scoring of high quality compositions. The above-mentioned variations in scores negatively affect the reliability, validity, and fairness of the judgments about a student's writing performance (Han, 2013; Huang, 2008, 2009, 2011).

As the above research literature has indicated, some research has investigated the scores assigned to different quality EFL compositions (Mendelsohn & Cumming, 1987; Cumming, 1985; 1990; Huang, et al., 2014), but not the link between the quality of the papers, the ratings and the raters' decision-making behaviors while rating compositions at these three levels. Further, among these studies very few of them have used think-aloud protocols to analyze rater behaviors, especially in rating different quality writing (Cumming, 1990), although some research has benefitted from it in examining rating process, validity and fairness issues (Connor-Linton, 1995; Cumming, 1990; Cumming, Kantor & Powers, 2001; Sakyi, 2000; Vaughan, 1991; Weigle, 1994). On the other hand, research literature has indicated that empirical studies that investigate rater variation should look closely at the rating process (Connor-Linton, 1995) because think-aloud protocol data provides the "richest evidence" on the raters' decision-making behaviors while rating compositions. Therefore, this study aims at bridging this research gap. The following main research question guided this study:

What is the variability in and reliability of the scores assigned by EFL instructors to different quality compositions and how do they behave while rating the compositions? Further, within the G-Theory framework, the following three specific research questions were asked:

a. Are there differences among the holistic scores of the three different qualities of EFL papers?
b. What are the sources of score variation contributing to the score variability in the holistic scores assigned to the different quality EFL papers?
c. Does the reliability (e.g., dependability coefficients for criterion-referenced score interpretations) of the holistic scores differ among the scores assigned to the different quality EFL papers?

Further, the data derived from the think-aloud protocols with the raters were used to answer the following additional research question:

How do the raters make decisions while marking different quality EFL papers holistically?

Methodology

The purpose of this study was to examine both the impact of writing quality (low-medium-high) on the variability in and reliability of EFL writing assessments in Turkey, and also rater behaviors using a mixed-methods research approach. Quantitatively, rating variability and reliability issues were examined employing the G-Theory approach. Qualitatively, think-aloud protocols were used with raters to explain the quantitative results further. The main research question was to determine "if there were any differences in the rating variability and reliability of the holistic writing scores assigned by the instructors to three different qualities of composition, and if the raters behaved differently across different qualities of papers while rating".

Context of the study

This study used writing quiz data, collected as part of the evaluation dimension of undergraduate classroom-based English examinations at the English Language and Literature Department of a state university in Turkey, where the medium of instruction is English. Generally, in the English Language and Literature Department, undergraduate classroom writing assignments and quizzes require undergraduate students to write an essay on one prompt in 45 minutes. Each time students are asked to write on a single topic that has been chosen for all students.

Obtaining Data

An instructor at the English Language and Literature Department of a state university provided the writing samples necessary for the analysis. Firstly, argumentative short compositions, written using a word processor computer program (e.g. Microsoft word), by EFL students who took English Writing Skills courses were selected. Secondly, the participant instructors were first informed briefly about the context of the study. Then, they were invited to participate in the study as volunteer participating raters. Five raters were randomly selected from among the volunteers to rate the essays. In addition, think-aloud data was collected from the raters about their rating processes. The purpose of the think-aloud protocols was to elicit how they arrived at holistic scoring decisions while rating very low, mid-range and very high quality compositions.

Selection of Writing Samples

The selection of the writing samples was undertaken as follows. Initially, all English instructors from three different universities were randomly allocated students' English writing samples, selected for data analysis from departmental impromptu writing examinations. The course teacher had not discussed the essay topics in class beforehand. Both tasks were argumentative and authentic in nature and the students were thought to be familiar with the content material used. The students were asked to respond to different argumentative writing tasks as an assignment.

Each instructor selected 10 argumentative essays written by nine undergraduate students outside of the class. These 30 papers were written by Turkish-speaking students as home assignments outside of the class and evaluated by the instructor as representing three different levels of quality (high, medium, and low) in order to maximize the differences among papers. A total of 30 papers were selected for this study.

Selection of Raters

The five participating raters were selected from volunteering lecturers with various teaching backgrounds, but who were at least studying for an MA degree in the field of interdisciplinary EFL and who had 1-3 years' experience in EFL teaching. They had similar teaching and assessing ESL/EFL writing experience. All the raters who participated in this experimental study are professionals in the field of interdisciplinary English language teaching and regular employees of three universities in Turkey.

These five raters were all graduates from different English Language Teaching or English Language and Literature departments in Turkey. They were Turkish native speakers of English. The ages of the raters ranged from 25 to 30. Their experiences in teaching EFL writing and marking EFL essays were similar.

Rating Scale

The instrument used in the study was a 10-point holistic scale, including five levels of score (see appendix A).

Training Raters to use the holistic scale in rating

The rating procedure for this study consisted of three phases. First of all, the names of the students were deleted from the papers and they were given unique codes to provide unbiased conditions for raters. Secondly, two hours of rater training was given to the participant raters about how to use the scoring rubric. Then raters gave grades for each composition using the holistic rubric.

Training Raters in How to Think-aloud

To train the raters to think-aloud, they were informed, in a two hour course, about what think-aloud is and about how the think-aloud procedure is handled. Then, the raters read some articles and book chapters about the think-aloud procedure. After that, they watched some sample videos about the procedure. Lastly, all of the raters made a sample think-aloud record, listened to the recordings and made comments about each other's. After the think-aloud procedure had been explained, each rater recorded their-rating procedure for 6 compositions (2 high, 2 medium, 2 low) without knowing the qualities of the papers while rating.

Quantitative Data analysis

Descriptive statistical analysis (mean and standard deviation) was conducted for the holistic scores of the low, medium, and high quality papers, respectively. The purpose of conducting descriptive statistical analysis was to obtain a general comparison of both the mean score and standard deviation differences among the papers of different qualities. Within the G-Theory framework, data were analyzed in the following three stages: 1) student nested within quality-by-rater (with paper quality fixed and all other facets random) mixed effects G-study; 2) student-by-rater random effects G-studies for low, medium, and high quality papers, respectively, and 3) calculation of G-coefficients (Huang & Foote, 2010).

Student nested within quality-by-rater mixed effects G-study

A student nested within quality -by-rater (s:q x r) mixed effects G-study analysis (with paper quality fixed and all other facets random) was conducted. The purpose of this G- study was to obtain variance component estimates for the six independent sources of variation: quality (q), student nested within quality (s:q), rater (r), quality-by-rater (q x r), student nested within quality-by-rater (s:q x r), and the residual (p:q x r).

Paper-by-rater random effects G-studies

Three separate paper-rater (p x r) random effects G-studies were conducted for low, medium, and high quality papers, respectively. The purpose of these G-studies was to obtain information for comparison among the low, medium, and high quality papers, in terms of score variability and reliability. It was hypothesized that there would be differences among these compositions of different qualities. With the implementation of these G-studies, the three independent sources of variation, namely, student (p), rater (r), and student-by-rater (p x r) for each quality level were obtained. Using the obtained variance components, G-coefficients for each quality level were then calculated in order to examine for reliability (cf. Huang, 2012).

Calculation of G-coefficients

Two different reliability coefficients (phi- and G- coefficients) related to decisions (the interpretation of the criterion-referenced level of scores and of the norm-referenced level of scores) can be calculated through G-Theory analysis (Shavelson & Webb, 1991). Based on the paper-by-method-by-rater (s x r) random effects G-studies results, the G-coefficient and phi-coefficient for each level of paper quality (low, medium, high) were calculated. The purpose of calculating these coefficients for each level of paper quality was to answer the second research question: Does the reliability of scores differ among essays of three different qualities? The computer program EduG was used for the G-studies. EduG is a computer program "based on the Analysis of Variance (ANOVA) and Generalizability Theory (G-Theory), and designed to carry out generalizability analysis" (EduG, 2015).

Qualitative Data Analysis

Think-aloud protocol analysis for this study was conducted in several steps, following Cumming, Kantor and Power's (2001) protocol analysis approach. First, protocols were transcribed by two volunteer teachers and revised to a standard set of simple transcription conventions (c.f. Cumming, Kantor & Powers, 2001). Second, the transcriptions were double checked to assure accuracy by the researcher of this study and another researcher. Third, Cumming, Kantor and Power's (2001) modified version of the schemes developed by Cumming (1990) and Sakyi (2000), (see Appendix B) was used to describe decision-making behaviors while rating EFL compositions. This analysis identified three general decision-making categories that the raters made during the ratings of compositions. These categories were a) "self-monitoring of one's own rating behaviors", b) "the composition's realization", and c) "rhetorical and ideational elements, or the composition's accuracy and fluency in the English language" (Cumming, et al., 2001, p.16).

Results

This section includes quantitative and qualitative data analysis regarding the study. Descriptive analysis for the study is presented first, followed by G-Theory analyses, and, finally, the think-aloud protocol analysis is presented.

Descriptive Statistical Results

Each of the papers was rated holistically by five independent raters on a ten-point rubric. Table 1 provides the descriptive statistics (i.e., the mean and standard deviation) for the scores of the very low, mid-range and very high quality papers used in the analysis.

Table 1. *Descriptive Statistics for very low, mid-range and very high quality papers*

#Papers	Very low level		Mid-range		Very high level	
	m	**sd**	**m**	**Sd**	**m**	**sd**
#1	5.50	1.93	6.30	1.92	8.60	0.42***
#2	5.20	1.30	6.50	2.74**	7.80	1.64
#3	4.80	1.26	6.20	2.20**	7.70	1.30
#4	6.30	1.92	5.80	1.30	8.20	0.97
#5	4.90	1.43	7.00	1.22	8.00	1.37
#6	3.60	0.96*	7.00	1.46	8.40	1.08
#7	5.50	1.66	6.00	1.70	8.10	0.89
#8	5.20	1.35	6.40	2.27**	8.00	0.71
#9	5.00	2.37**	6.90	1.43	6.20	1.04
#10	5.80	2.17**	6.80	1.35	6.90	0.42***

Comparing the results of the very low, mid-range, and very high quality papers, the following three observations were made. First, the standard deviations for both the very low and mid-range quality papers are larger than 1.0, except for paper #6 in the low quality category; and, in the high quality category, except for papers *1#, 4#, #7, #8 and #10, indicating that raters scored the very high quality papers more consistently compared to the other two quality levels. Further, papers **#9 #10 in the low category and papers **#2 # 3 and # 8 in the middle category have a large standard deviation of over 2, whereas, for the high quality papers, students ***#1 and #10 have standard deviations smaller than 0.5, indicating that the raters mostly deviated while giving scores to very low level and mid-range compositions, but that they were more consistent while rating very high level compositions.

G-theory Analyses

Student Nested within Quality-by-Rater Mixed Effects G-study Results

A student nested within quality-by-rater (s:q x r) mixed effects G-study analysis was conducted to obtain variance component estimates for the six independent sources of variation. The results are presented in Table 2.

Table 2. *Variance components for random effects SXR:Q design*

Source of variability	Df	σ^2	%
S:Q	27	0.25611	6.0
R	4	1.05621	24.8
Q	2	1.58343	37.2
RQ	8	0.38472	9.0
SRQ	108	0.97241	22.9
Total	149		100%

Table 2 reveals that the largest variance component (37.2% of the total variance) was attributable to the quality of the papers, in other words that there were considerable differences between compositions in terms of the standard of writing performance. The second largest variance component (4.39% of the total variance) was attributable to rater (r), in other words the raters differed in terms of the severity or leniency of their rating. The third largest component (22.9% of the total variance) was the residual variability arising from the interaction of the raters, the quality of the compositions and various unelucidated sources of error, whether systematic or unsystematic (Huang, 2007, 2008, 2012). The fourth largest variance component (9 % of the total variance) was attributable to quality-by-rater (qr), in other words there were large differences between raters in their ratings of papers of different quality. The fifth largest variance component (6 % of the total variance) was attributable to student nested within quality (s:q), in other words that scoring of students was very

different within each level of quality.

Table 3. *Variance components for random effects s x r G-study design (Low quality papers)*

Source	df	σ^2	%
S	9	0.29417	9.3
R	4	1.77750	56.4
SR	36	1.08250	34.3
Total	49		100%

Table 3 gives the student-by-rater results of the random effects G-studies for low quality papers. It reveals that the largest variance component (56.4% of the total variance) was attributable to rater (r), in other words that there was a considerable difference between raters in respect of the leniency with which they scored the ten low quality EFL papers. The second largest variance component (34.3% of the total variance) was the residual variability arising from the interaction of the raters, the papers and various unelucidated sources of error, whether systematic or unsystematic (Huang, 2007, 2008, 2012). The lowest variance component (9.3 % of the total variance) was attributable to the object of measurement, student (s), which reveals that the ten selected low quality EFL papers did not differ markedly in respect of quality.

Table 4. *Variance components for random effects s x r G-study design (middle quality papers)*

Source	Df	σ^2	%
S	9	0.07444	2.2
R	4	2.05028	61.3
SR	36	1.21722	36.4
Total	49		100%

Table 4 gives the student-by-rater results of the random effects G-studies for student papers of medium quality. These are similar to those in Table 3. They reveal that the largest variance component (61.3% of the total variance) was attributable to rater (r), in other words that there was a considerable difference between raters in respect of the leniency with which they scored the ten medium quality EFL papers. The second largest variance component (36.4% of the total variance) was the residual variability arising from the interaction of the raters, the papers and various unelucidated sources of error, whether systematic or unsystematic (Huang, 2007, 2008, 2012). The lowest variance component (2.2 % of the total variance) was attributable to the object of measurement, student (s), which indicates that the ten selected middle quality EFL papers did not differ markedly in respect of quality.

Table 5. *Variance components for random effects s x r G-study design (high quality papers)*

Source	df	σ^2	%
S	9	0.39972	26.4
R	4	0.49500	32.7
SR	36	0.61750	40.8
Total	49		100%

Table 5 gives the student-by-rater results of the random effects G-studies for student papers of high quality. They reveal that the largest variance component (40.8% of the total variance) was attributable to the residual variability arising from the interaction of the raters, the papers and various unelucidated sources of error, whether systematic or unsystematic (Huang, 2007, 2008, 2012). The second largest variance component (32.7% of the total variance) was attributable to rater (r), which indicates that there was a considerable difference between raters in respect of the leniency with which they scored the ten high quality EFL papers. The lowest variance component (26.4 % of the total

variance) was attributable to the object of measurement, student (s), which indicates that the ten selected high quality EFL papers did not differ markedly in respect of quality

Using the student-by-rater random effects G-studies variance component results, the G-coefficients for each quality were calculated and presented in Table 6.

Table 6. Dependability coefficients for ratings of low quality papers

	papers	raters	Phi coefficient	g-coefficiency
	30	5	.87	.79
High quality	10	5	.76	.64
Middle quality	10	5	.23	.10
Low quality	10	5	.58	.34

As shown in Table 6, the phi-coefficient and G-coefficient obtained for the ratings of high quality papers for the current five-rater scenario (.87 and .79 respectively) were higher than the coefficients obtained for the mid-range and low quality compositions, while the lowest coefficients were obtained for the ratings of middle quality papers (.10 and .23, respectively). Further, the results show that both the phi-coefficient and the G-coefficient obtained for the ratings of the low quality papers were more than two times higher than the ratings of the middle quality papers (.58 and .34, respectively).

Qualitative Data Analysis

The think-aloud protocols by 5 raters recorded for their rating procedure regarding the 6 compositions (2 high, 2 mid, 2 low) included different numbers of protocols. Table 7 shows the number of protocols and percentages for the three categories (self-monitoring, rhetorical and ideational, and language focus) in respect of the three qualities of composition (very high, mid-range and very low).

Table 7. *Number and percentage of protocols for decision-making behaviors involving the three types of focus, for essays rated very low, mid range or very high*

Very low level compositions						
	Self-monitoring focus		Rhetorical and ideational focus		Language Focus	
RATERS	n	%	n	%	n	%
R1	7	5.47	8	6.25	17	13.28
R2	2	1.56	10	7.81	16	12.50
R3	4	3.13	2	1.56	8	6.25
R4	5	3.91	6	4.69	13	10.16
R5	7	5.47	13	10.16	10	7.81
Total:128	***25***	***19.53***	***39***	***30.47***	***64***	***50***
Mid-range level compositions						
	Self-monitoring		Rhetorical		Language Focus	
	n	%	n	%	n	%
R1	3	2.10	7	4.90	12	8.39
R2	3	2.10	11	7.69	16	11.19
R3	5	3.50	1	0.70	7	4.90
R4	4	2.80	6	4.20	18	12.59
R5	5	3.50	6	4.20	39	27.27
Total:143	***20***	***13.99***	***31***	***21.68***	***92***	***64.34***
Very high level compositions						
	Self-monitoring		Rhetorical		Language Focus	
	n	%	n	%	n	%
R1	5	3.50	7	4.90	15	10.49
R2	6	4.20	9	6.29	16	11.19
R3	3	2.10	2	1.40	8	5.59
R4	3	2.10	5	3.50	20	13.99
R5	4	2.80	12	8.39	28	19.58
Total:143	***21***	***14.69***	***35***	***24.48***	***87***	***60.84***

Table 7 presents the number and frequency of protocols for all decision-making behaviors, including self-monitoring, ideational and rhetorical, and language for essays rated very low, mid range or very high by the EFL composition teachers. The results indicated that, firstly, even though the raters attended differently to different aspects of the three level compositions, the raters did more self-monitoring of their assessment behaviors for the very low level compositions (19.53 %). Yet, for mid-range and very high-level compositions the raters behaved similarly (13.99 % and 14.69 %, respectively). Further, the raters seem to have varied much in self-monitoring behavior while rating very low-level compositions. Next, the raters attended more to rhetoric and ideas when they rated very low-level compositions (30.47%) compared to compositions at the other levels (21.68 % and 24.48 % respectively). Further, the raters seem to be more consistent in this behavior while rating very high level compositions. Finally, when rating mid-range and very high-level compositions, the raters paid relatively more attention to language features (64.34 % and 60.84% respectively).

Discussion and conclusion

The first research question attempted to examine if there was any score variation among the scores assigned by the raters to the EFL compositions of different qualities. The descriptive results indicated that the raters mostly deviated while giving scores to very low level and mid-range compositions, but that they were more consistent while rating very high-level compositions. This is because “essays which fall in the midrange are often most difficult for readers to assess since they usually contain characteristics of high and low levels of writings” (Elbow, 1993 cited in Russikoff,

1995, p.2; Hamp-Lyons, 1991). For example, Mendelsohn and Cumming (1987) and Huang et al., (2014) found that raters differed on the ratings of middle quality papers.

The second research question examined the sources of score variation contributing to the score variability of the holistic scores assigned to the different quality EFL papers. The G-Theory analyses showed that there was a large difference in writing performance that could be attributed to the qualities of essays (37.2%). Further, raters did differ considerably from one another in terms of leniency in marking the 10 very low and 10 mid-range quality EFL papers (56.4 % and 61.3% respectively), whereas raters (r) (32.7% of the total variance) differed less from one another in terms of leniency in marking the 10 high quality ESOL papers. This G-Theory analysis confirmed the descriptive results, as there was more score consistency in the ratings of high quality papers. These results were consistent with those of a previous study (Huang, et al., 2014). Further, Esfandiari and Myford (2013) indicated that variability in writing scores stems from the type of assessment, and that teacher assessors could be more severe than self-assessors and peer assessors. Yet, Lim (2009) argued that severity differences did not produce any significant variation in writing scores at the end. In this context, Huang (2008) has proposed that it is essential to prevent discrepancies in rater severity and leniency when a great number of raters participate in the writing assessment process.

The third research question aimed at investigating the reliability of relative and absolute (phi-coefficient and G-coefficient) score differences assigned to compositions at the three levels of quality. The reliability coefficients for the ratings of high quality papers (g: .87 and phi: .79, respectively) were higher than the coefficients obtained for the mid-range and low quality compositions, with the lowest coefficients obtained for the ratings of middle quality papers (.10 and .23 respectively). Further, the coefficients obtained for the ratings of the low quality papers were more than two times higher than for the ratings of the middle quality papers (.58 and .34 respectively). This result indicated that more reliable ratings could be obtained for high quality papers. This result is in parallel with a study by Huang et al. (2014) which found that raters were most consistent in rating high quality papers. However, it was also found that the scoring method (holistic or analytic) greatly impacted upon the scoring of high quality EFL papers, but that the scoring method did not impact upon the scores for low quality papers.

Finally, the last research question was aimed at examining raters' decision making behaviors (self-monitoring, ideational and rhetorical, or language) while rating the three different quality compositions. The raters attended differently to different aspects of these three levels of composition. Specifically, while the raters did more self-monitoring of their assessment behaviors on the very low-level compositions (19.53 %), they behaved similarly for mid-range and very high-level compositions (13.99 % and 14.69 %, respectively). Further, there was much variation in self-monitoring behavior while rating very low-level compositions. Next, the raters attended more to rhetoric and ideas when they rated very low-level compositions (30.47%) compared to mid-range and high quality compositions (21.68 % and 24.48%, respectively). Further, the raters seemed to be more consistent in this behavior while rating very high level compositions. Finally, when rating mid-range and very high-level compositions, the raters seemed to devote relatively more attention to language matters (64.34% and 60.84%, respectively). There is a discrepancy between the results of this study and previous research in this respect: one study found that ESL professors gave more weight to rhetorical organization (Mendelsohn & Cumming, 1987), and some other studies have indicated that raters mostly give lower scores to compositions that have poor linguistic features such as lexicon and simple sentence structures (Engber, 1995; Russikoff, 1995; Song & Caruso, 1996; Vaughan, 1991).

There are two limitations that need to be acknowledged regarding this study. Firstly, the large residual variance component for low, mid-range, and high quality EFL papers (34.3 %, 36.4% and 40.8% respectively) in the G-studies indicates that other facets might have attributed to the score variance. Large residual effects may stem from hidden facets (Brennan, 2001). "The variance of the hidden facets is included in the residual variance, thus leading to a larger residual than when the facet is explicitly considered" (Huang et al., 2014, p.144). Secondly, this study included a single task type. However, available research has found that task types affect the variability and reliability of composition scores (Huang, 2008; Lee & Kantor, 2005).

In conclusion, the results of this study have proved that the link between paper quality and rating scores is variable. Moreover, this variation directly affects fairness issues. It can be inferred that teachers of writing and assessment professionals should receive comprehensive training regarding how to assess different qualities of composition (Huang & Foote, 2010; Huang et al., 2014). Finally, further research could investigate the link between task types, rater experience and scores assigned to different qualities of composition.

Acknowledgement

I would like to thank Gamze Tekin who helped me in double-checking and coding the data, Elif Tokdemir-Demirel for providing the compositions and finally Catherine Akça for proof reading the paper.

References

Alaei, M. M., Ahmadi, M., & Zadeh, N. S. (2014). The impact of rater's personality traits on holistic and analytic scores: Does genre make any difference too?. *Procedia-Social and Behavioral Sciences*, *98*, 1240-1248. http://dx.doi.org/10.1016/j.sbspro.2014.03.539

Algina J. & Swaminathan H. (2015). Psychometrics: Classical test theory. *International Encyclopedia of the Social & Behavioral Sciences*, 423-430. http://dx.doi.org/10.1016/B978-0-08-097086-8.42070-2

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language.*Journal of Second Language Writing, 18*, 191-208. http://dx.doi.org/10.1016/j.jslw.2009.05.003

Bacha, N., (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us?*System,29*, 371-383. http://dx.doi.org/10.1016/S0346-251X(01)00025-2

Baker, B. A. B. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing,15*,133–153. http://dx.doi.org/10.1016/j.asw.2010.06.002

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*, 86–107. http://dx.doi.org/10.1016/j.asw.2007.07.001

Breland, H., Bridgeman, B., & Fowles, M. (1999). Writing assessment in admission to higher education: Review and framework (GRE Board Report No. 96-12R). Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-99-03-Breland.pdf on October 26, 2015

Brennan R. L. (2010). Generalizability theory. *International Encyclopedia of Education*, 61-68. http://dx.doi.org/10.1016/B978-0-08-044894-7.00246-3

Brennan R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1-21. http://dx.doi.org/10.1080/08957347.2011.532417

Brennan, R. L. (2001). *Statistics for social science and public policy: Generalizability theory*. New York: Spr inger -Verlag.

Briesch A. M., Swaminathan H., Welsh M. & Chafouleas S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology, 52* (1), 13-35. http://dx.doi.org/10.1016/j.jsp.2013.11.008

Chang, Y. (2002). EFL teachers' responses to L2 writing. Reports Research (143). Retrieved from http://files.eric.ed.gov/fulltext/ED465283.pdf on March 23, 2015

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29*(4), 762-765. http://dx.doi.org/10.2307/3588174

Cumming, A. (1985). *Responding to the writing of ESL students*. In Patterns of development. A. Pare

& M. Maguire (Eds.). Ottawa: Canadian Council of Teachers of English.

Cumming, A. (1990). Expertise in evaluating second language composition. *Language Testing, 7*, 31-51. http://dx.doi.org/10.1177/026553229000700104

Cumming, A., Kantor, R., & Powers, D.E. (2001). Scoring TOEFL Essays and TOEFL 2000 Prototype Writing Tasks: An Investigation Into Raters' Decision Making and development of a Preliminary Analytic Framework (TOEFL Monograph Series Rep. No. 22). Princeton, NJ: Educational Testing Service.

EDUG (2015). Edug. Retrieved from http://www.irdp.ch/edumetrie/eduGeng.htm on October 15, 2015.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*, 139-155. http://dx.doi.org/10.1016/1060-3743(95)90004-7

Eryaman, M. Y. (2008). Writing, method and hermeneutics: Towards an existential pedagogy. *Elementary Education Online, 7*(1), 2-14.

Esfandiari, E., & Myford, C. M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing,18*(2), 111–131. http://dx.doi.org/10.1016/j.asw.2012.12.002

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. Assessing Writing, 21, 56–73. http://dx.doi.org/10.1016/j.asw.2014.03.002

Green, J. M. & Oxford R. (1995). A closer look at learning strategies, L2 proficiency, and gender. *Tesol Quarterly, 29*(2), 261-418. http://dx.doi.org/10.2307/3587625

Han, T. (2013). The impact of rating methods and rater training on the variability and reliability of EFL students' classroom-based writing assessments in Turkish universities: An investigation of problems and solutions. Atatürk university, Erzurum, Turkey.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of Classical test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hamp-Lyons, L. (1991). Scoring Procedures for ESL Contexts. L. Hamp-Lyons (Ed.). *Assessing Second Language Writing in Academic Contexts* (pp. 241-277). Norwood, NJ: Ablex

Heaton J. B. (2003). *Writing English language tests*. USA: Longman.

Huang, J. (2007). *Examining the Fairness of Rating ESL Students' Writing on Large-Scale Assessments* (Unpublished Doctoral Dissertation). Canada: Queen's University.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? – A generalizability theory approach. *Assessing Writing, 13*(3), 201-218. http://dx.doi.org/10.1016/j.asw.2008.10.002

Huang, J. (2009). Factors Affecting the Assessment of ESL Students' Writing. *International Journal of Applied Educational Studies, 5*(1), 1-17.

Huang, J. (2011). Generalizability Theory As Evidence of Concerns about Fairness in Large-Scale ESL Writing Assessments. *TESOL Journal, 2*(4), 423-443. http://dx.doi.org/10.5054/tj.2011.269751

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing. *Assessing Writing, 17*(3), 123-139. http://dx.doi.org/10.1016/j.asw.2011.12.003

Huang, J., & Foote, C.J. (2010). Grading Between Lines: What Really Impacts Professors' Holistic Evaluation of ESL Graduate Student Writing? *Language Assessment Quarterly, 7*(3), 219-233.

Huang J., Han, T., Tavano, H., & Hairston, L (2014). Using generalizability theory to examine the impact of essay quality on ESOL writing assessment- A Turkish case study. China- US Education, 3, 3-20.

Knoch, U., (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from?. *Assessing Writing,16*(2), 81-96. http://dx.doi.org/10.1016/j.asw.2011.02.003

Kobrin, J. L., Deng, H., & Shaw, E. J. (2011). The association between sat prompt characteristics, response features, and essay scores. *Assessing Writing, 16*,154–169. http://dx.doi.org/10.1016/j.asw.2011.01.001

Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL Writing Test Scores: Evaluating Prototype Tasks and Alternative Rating Schemes* (TOEFL Monograph No. MS-31). Princeton, NJ: ETS.

Lim, G. S. (2009). Prompt and rater effect in second language writing performance assesment. (Doctoral dissertation, The University of Michigan). Retrieved from http://deepblue.lib.umich.edu on March 23, 2015

Matt G. E. & Sklar M. (2015). Generalizability theory. *International Encyclopedia of the Social & Behavioral Sciences*, 9, 834-838. http://dx.doi.org/10.1016/B978-0-08-097086-8.44027-4

Mendelsohn, D., & Cumming, A. (1987). Professors' Ratings of Language Use and Rhetorical Organization in ESL Compositions. *TESL Canada Journal, 5*, 9-26.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing,15*(1) , 18-39. http://dx.doi.org/10.1016/j.asw.2010.01.003

Rindskopf D., (2015). Reliability: Measurement. *International Encyclopedia of the Social & Behavioral Sciences, 20*, 248-252. http://dx.doi.org/10.1016/B978-0-08-097086-8.44050-X

Russikoff, K. A. (1995). *A Comparison of Writing Criteria: Any Differences?*, [Proceeding]. *Paper Presented at the Annual Meeting of the Teachers of English to Speakers of Other languages*, Long Beach: CA.

Saeidi, M., & Semiyari, S. R. (2011). The impact of rating methods and task types on EFL learners' writing scores. *Journal of English Studies*, 1(4), 59-68.Retrieved from http://www.sid.ir on March 25, 2015

Sakyi, A. (2000). Validation of holistic rating for ESL writing assessment: How raters evaluate ESL compositions. In A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129-152). Cambridge: Cambridge University Press.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability Theory: A Premier*. Newbury Park, CA: Sage.

Shi, L. (2001). Native- and Nonnative-Speaking EFL Teachers' Evaluation of Chinese Students' English Writing. *Language Testing, 18*(3), 303-325.

Song, B., & Caruso, I. (1996). Do English and ESL Faculty Differ in Evaluating the Essays of Native English-Speaking, and ESL Students? *Journal of Second Language Writing, 5*, 163-182. http://dx.doi.org/10.1016/S1060-3743(96)90023-5

Steyer R. (2001). Classical (psychometric) test theory. *International Encyclopedia of the Social & Behavioral Sciences*, 3, 1955-1962. http://dx.doi.org/10.1016/B978-0-08-097086-8.44006-7

Vaughan, C. (1991). Holistic Assessment: What Goes on in the Raters' Minds?, L. Hamp-Lyons (Ed.).*Assessing Second Language Writing in Academic Contexts* (pp. 111-126). Norwood, NJ: Ablex.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*, 197-223. http://dx.doi.org/10.1177/026553229401100206

Weigle, S. C. (2008). *Assessing writing*. UK: Cambridge University Press

Willingham, H. H., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Lawrence Erlbaum.

Appendices

Appendix A: 10-point Holistic Scale

Score	Criteria
8.5-10	Natural English and no direct translation of idioms and phrases from Turkish. Excellent choice of vocabulary. Complete knowledge of syntax and morphology. Appropriate use of articles and prepositions. Good spelling, punctuation and capitalization. Topic is clearly stated. All parts of the text have excellent unity and coherence.
6.5-8	Sufficient naturalness of English and few collections of simple sentences and direct translations of idioms from Turkish. Good vocabulary choice. Extensive knowledge of syntax and morphology. Few random uses of articles and prepositions. A few spelling, punctuation and capitalization errors. Topic is rather clear. All parts of the text have good unity and coherence.
5-6	Lack of naturalness of English and not many direct translations of idioms and phrases from Turkish. Average vocabulary choice. Moderate knowledge of syntax and morphology. Some inappropriate use of articles and prepositions. There are several spelling, punctuation and capitalization errors. The topic is stated but it is not clear. All parts of the text have average level of unity and coherence.
2.5-4.5	Poor and informal English and frequent direct translations of idioms and phrases from Turkish. Weak choice of vocabulary. Limited knowledge of syntax and morphology. Serious errors in articles and prepositions. Spelling, punctuation and capitalization errors are common. The topic is unrelated. All parts of the text have poor level of unity and coherence.
0-2	Insufficient naturalness of English and many direct translations of idioms and phrases from Turkish. Very week vocabulary choice. No evidence of knowledge of syntax and morphology. Nearly all the articles and prepositions are used wrong. Many spelling, punctuation and capitalization errors. Topic is missing. The text has nearly no unity and incoherent.

Appendix B: Decision-Making Behaviors While Rating ESL Compositions

Self-monitoring focus	Task fulfillment: rhetorical and ideational focus	Language Focus
Interpretation strategies		
* scan whole text * envision situation of writer * focus self on task rubric	* interpret ambiguous phrases * discern rhetorical structure * summarize propositions	* classify error types * “edit” phrases for interpretation
Judgement Strategies		
* establish personal response * define and revise own criteria * compare with other compositions or anchor papers * distinguish interactions between categories * summarize judgments collectively * articulate scoring decision	* assess total output * assess relevance * assess coherence * assess interest * identify redundancies * assess topic development * assess helpfulness in guiding reader * rate content and organization overall	* establish level of comprehensibility * establish error values * establish error frequency * establish command of lexis * establish command of syntax and morphology * establish command of spelling and punctuation * rate language overall